# PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2003-256438
(43)Date of publication of application : 12.09.2003

(51)Int.Cl. G06F 17/30

(21)Application number : 2002-056129
(22)Date of filing : 01.03.2002

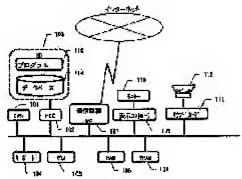
(71)Applicant : CANON INC
(72)Inventor : TAMAI SHUNICHI

## (54) MULTIMEDIA DATABASE SYSTEM, STORAGE MEDIUM AND PROGRAM

(57)Abstract:

PROBLEM TO BE SOLVED: To provide a database, capable of facilitating registration and having a correspondence with copyright.

SOLUTION: A multimedia database system includes an automatically document classifying means for classifying Web page into categories according to a user's intention, a means for extracting a specified text having relation with the multimedia data from the Web page, a means for retaining a database capable of registering and managing the category information obtained from the multimedia data and the automatically document classifying means and the specified text, a means for extracting the information related to copyright on the multimedia from the Web page and a means for registering and managing the extracted information.



* NOTICES *



1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

## DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Field of the Invention]This invention peruses the web page on the Internet, and relates to the multimedia database system which carries out registration management of the multimedia data on a web page.
[0002]
[Description of the Prior Art]Into the website of the Internet, accumulate the huge information on the Internet called a search site, and a database is built, When a user inputs a keyword there, there is a site which displays URL of a website, a link, an abstract, etc. with the information about the keyword.
[0003]In order to build the database at such a site, there are two methods, one is a means called a robot type, and another is a means called a directory type.
[0004]A robot type puts in a database the information on the web page which automatic-patrolled the server in the world periodically and collected them. Since it goes round automatically, as for an advantage, there are many web pages registered into a database.
[0005]As for a directory type, a database is built when the author of a web page or the administrator of a search site registers. Since a specific person registers, there are few web pages registered into a database, but the reliability is high.
[0006]The information registered into the database of these search sites is mainly text data. However, the art of building a multimedia database in part using the multimedia data on a web page is also proposed. For example, it is JP,2000-076300,A etc.
[0007]
[Problem(s) to be Solved by the Invention]However, even if it builds multimedia database system using the above-mentioned conventional data base construction system, in a robot type case, it is usually unreliable, and when it actually searches, aim has achieved many

results of a gap plentifully in addition to the expected information.
[0008]The built multimedia database was unsuitable, when it became large-scale, mass memory storage was needed and individual multimedia database system was generally considered. A reliable directory type has the problem that people's work keeps once in many as mentioned above at the time of registration.
[0009]It is not allowed to use the downloaded multimedia data without a right holder's consent except for cases, such as enjoying oneself individually, from a viewpoint of copyright protection to the multimedia data on the page referred to. So, registration management of the information about the owner of a copyright corresponding to each multimedia data needs to be carried out to multimedia database system.
[0010]
[Means for Solving the Problem]This invention is characterized by the 1st invention concerning this application comprising the following, in order to solve an aforementioned problem.
It connects with a web server on networks, such as the Internet, and is a browsing means which can download web page information.
A document automatic-classification means to classify said web page into a category along with a user's intention.
A means to extract multimedia data information from said web page information, and to download said multimedia data.
A means to extract a specific text with said multimedia data and relevance from said web page, A means to hold a database which can carry out registration management of said multimedia data, category information acquired from said document automatic-classification means, and said specific text, A means to extract information relevant to copyright about first half multimedia data from said web page, a means which carries out registration management of the extracted information, and a means to operate said database.

[0011]This invention is characterized by the 2nd invention concerning this application comprising the following, in order to solve an aforementioned problem.
Said browsing means is a URL (Uniform Resource Locator) input part.
A data download part which downloads HTML (Hypertext Mark up Language) source data from a worldwide web server.
A data analysis part which analyzes downloaded HTML source data.
A display-image generation part which creates a display image of a web page according to an analysis result, a fill out form input part for inputting data to a data entry column detected in the analysis of page data, and a hyper-ring extraction part which extracts hyperlink information from page data.

[0012]In order to solve an aforementioned problem, the 3rd invention concerning this application, Said multimedia database has a means to register category information, Information on said category by which the web page classification was carried out is used by said document automatic-classification means as category information at the time of registering said multimedia data into said multimedia database.
[0013]In order to solve an aforementioned problem, the 4th invention concerning this application is provided with said multimedia data, a means to extract a specific text with relevance, and a means to input said specific text into said multimedia database, from said web page.
[0014]In order to solve an aforementioned problem, the 5th invention concerning this application, It has said multimedia data, a means to extract two or more specific texts with relevance, and a means by which a user chooses said specific text registered into said multimedia database from said two or more extracted specific texts, from said web page.
[0015]In order to solve an aforementioned problem, the 6th invention concerning this application is provided with a means to extract information relevant to copyright about said multimedia data from said web page, and a means to register information relevant to said copyright into said multimedia database.
[0016]In order to solve an aforementioned problem, the 7th invention concerning this application, A means to extract two or more information relevant to copyright about said multimedia data from said web page, It has a means by which a user chooses information relevant to said copyright registered into said multimedia database from information relevant to said two or more extracted copyrights.
[0017]
[Embodiment of the Invention][Example 1] This example is one function of web browser application, and this explanation explains the multimedia database add function in connection with this invention of web browser application.
[0018]The lineblock diagram of this example is shown in drawing 1. 101 controls whole this example by CPU. 102 controls the data program in the hard disk (HD) of 103 by a hard disk controller (HDC). In the hard disk, the program 113 of web browser application with the function in connection with this invention and the database 114 are stored. 104 takes out a keyboard and 105 takes out directions of a program start etc. with pointing devices (PD), such as a mouse and a digitizer. 106 stores a program and data by RAM. 107 exchanges with the external Internet with a communication line interface (I/F). 109 performs control outputted to the monitor 110 by making into a video signal picture image data stored in VRAM108 with the display controller. 111 is a sound card and it is possible to output the voice data of multimedia data through the loudspeaker 112.
[0019]Drawing 2 is a block diagram of web browser application with the function in connection

with this invention. However, the portion without regards to this invention is simplified. In drawing 2, the main processing part 201 is a portion which performs control by the whole web browser application, and if URL is inputted by the keyboard 104 or the mouse 105 from a user, the information will be sent to the main processing part 201 via the GUI control section 202. Via the communications department 203, the Internet is accessed, and the main processing part 201 acquires the HTML file in URL inputted by the user from communication line I/F107, and stores it in the hard disk 103. The acquired HTML file is analyzed by the HTML analyzing parts 204, and the screen display information acquired as a result is expressed as the monitor 110 via the display controller 109 from the WEPEJI indicator 205.

[0020]Although an HTML file is acquired from the Internet, it is analyzed by HTML analyzing parts and the web page is displayed on the monitor in this example, The form of the data acquired by having a means to analyze the data for displaying homepages other than HTML analyzing parts is a thing that what is necessary is just data for not limiting to an HTML file and displaying a web page.

[0021]If there are directions of implementation of the multimedia database add function in connection with this invention with the keyboard 104 or the mouse 105 from a user, the directions will be sent to the main processing part 201 via the GUI control section 202. The main processing part 201 which received the directions starts multimedia database registration. The multimedia database registration method is explained in detail below using drawing 3 which explains the flow of processing of multimedia database registration in addition to drawing 2.

[0022]In drawing 3, the main processing part 201 memorizes the present time, URL of the HTML file, and the copyright information (a 212-page copyright information extraction part extracts) of the page as a variable to RAM106 by S0301. A HTML document is made into an example and the acquisition method of the copyright information in this example is explained.

[0023]From the last paragraph of page data, 1. "copyright", This sentence is made into owner-of-a-copyright information when the sentence containing character strings, such as "©", exists (a HTML tag is deleted), When 2. made into author contact information was also carried out and the e-mail address which continues after it when "mailto" exists similarly is not able to be acquired, When 3. for which it asks like one is also carried out and owner-of-a-copyright information is not able to acquire 1 and 2 from the inside of the head paragraph of page data, and a header, URL of the page is made into owner-of-a-copyright information, and data is emptied when author contact information is not able to be acquired. [0024]Here, although the case of the HTML document was described, in the case of other forms, it did not necessarily restrict at a described method, and it will be satisfactory if it is the method that copyright information can be acquired.

[0025]Next, in S0302, the main processing part 201 transmits the HTML file data stored in

HD103 to the sampling-of-text part 211 of drawing 2, and extracts (the thing except data parts other than the structure of a text, the portion showing expression, a link, and a document, etc.) only a text part here. For example, in the case of the contents written in HTML, processing which removes the tag data of HTML is performed. In HTML, since the tag was surrounded by "<" and ">", it was simply realized by eliminating the text of the inside surrounded by "<" and ">" and "<", and ">." The form of the data which carries out sampling of text is not limited to an HTML file, and what is necessary is just to perform sampling of text corresponding to each data format.

[0026]Like the following, in this example, a HTML document is made into an example and explained. Categorization is performed in the document automatic-classification part 206 to the text data of the web page obtained in the sampling-of-text part 211. About the technique of the categorization of S0303 in the document automatic-classification part 206, by JP,8-221429,A etc., since it is publicly known, explanation is already omitted here.

[0027]Next, the HTML tag for displaying the link or multimedia data from the head of page data (here HTML file) to multimedia data in the drawing 2 multimedia data extraction part 208 is investigated, and it goes by discernment of S0304 multimedia data. This pays its attention to the substance (file) name shown by URL following those tag attributes, when character strings, such as "A HREF" (refer to external file), "IMGSRC" (in-line image), and "ENBED" (external program etc.), are detected in a tag. In this example, it was judged simply whether it was multimedia data in the extension of this substance (file) name. For example, if it is still picture information, "gif", "jpg", "bmp", etc. are dynamic image data and "qt", "mov", "avi", etc. are voice data, they are "wav", "au", "aiff", etc.

[0028]Thus, URL in which multimedia data exists is acquired and multimedia data is acquired from communication line I/F107 from the Internet via the communications department 203 using the URL.

[0029]Next, in S0305, the display name expressing the acquired multimedia data is acquired from a HTML source code. It is among a code about the method of acquiring, and is shown below taking the case of the case where an in-line image is described.

[0030]1. If the ALT attribute as used in the field of a HTML tag is shown in the HTML tag for displaying the multimedia data, let the text data shown in the ALT attribute be a display name. Example: <blue IMG solvent refined coal = "aoiyama.gif" ALT = "mountain" ><BR> In the above-mentioned case, let "the blue mountain" in the ALT attribute in an IMG tag be a display name.

[0031]If the ALT attribute is not shown by 2.1., let the word (compound) which is just before the tag be a display name. Whether the sentence which is just before the tag is a word (compound). From the document analyzing information (since it is publicly known, detailed explanation is omitted) concerning the part of speech etc. of the word obtained during the

analysis of text automatic classification by S0303. Regarding it as a word (compound), when the text which is just before the tag is a text ended by a noun, the text ended except a noun does not regard it as a word (compound).
Example: When you are the mountain <BR> <IMG solvent refined coal ="kami-no-yama.gif"> above in which God lives, let the text data "mountain in which God lives" which comes out just before an IMG tag be a display name. However, in the following cases, it is not made into a display name.
As a mountain in which God lives, this mountain is collecting people's worship. <BR><IMG SRC = "kami-no-yama.gif"><BR> [0032]If the ALT attribute is not shown by 3.1., let the word (compound) immediately after the tag be a display name. Whether the sentence immediately after the tag is a word. Regarding it as a word (compound) from the document analyzing information (since it is publicly known, detailed explanation is omitted) acquired during the analysis of text automatic classification by S0303, when the text immediately after the tag is a text ended by a noun, the text ended except a noun does not regard it as a word (compound).
Example: When you are the mountain <BR> above in which <IMG solvent refined coal ="kami-no-yama.gif" <BR>> God lives, let the text data "mountain in which God lives" which comes out first be a display name after an IMG tag. However, in the following cases, it is not made into a display name.
<IMG solvent refined coal ="kami-no-yama.gif"> -- <BR <BR>> -- this mountain is collecting people's worship as a mountain in which God lives. <BR>[0033]4. When you do not fulfill the conditions of above-mentioned 1.2.3., let what removed the extension from the file name of multimedia data in URL which solvent refined coal shows with an IMG tag be a display name.
example: -- in IMG solvent refined coal ="http://www.URL.jp/hierarchy1/hierarchy2 / <momiji.gif" <BR>> above, By URL in the solvent-refined-coal attribute in an IMG tag, "http://" and a server name "www.URL.jp", Let "momiji" which removed the extension by the file name "momiji.gif" which removed a hierarchy "hierarchy1/" and "hierarchy2/" be a display name.
[0034]Next, by SO306, the drawing 2 data copyright information extraction part 213 acquires data copyright information. This method was performed as follows in this example.
[0035]The paragraph before and behind the HTML tag for displaying multimedia data is searched, and the existence of copyright information is judged. In this example, when the sentence containing character strings, such as "copyright", "©", and "registerd", existed, this sentence was made into owner-of-a-copyright information, and when "mailto" existed similarly, the e-mail address which continues after it was made into author contact information. However, when information does not exist in the searched paragraph, the copyright information acquired by SO301 is diverted.
[0036]Although the described method was adopted in this example, it does not restrict to this method, and if it is the method that copyright information can be acquired, it will be satisfactory

in any way. For example, many researches which embed copyright information etc. in the state of digital watermarking to the inside of a multimedia data body for the purpose, such as copyright protection and prevention from a data alteration, have come to be seen in recent years. If the program module which can read such electronic watermark information can be prepared, the detailed effective information about copyright can be acquired.

[0037]Next, URL of the target multimedia data is changed into a regular expression (SO307). since it is described using the relative describing method to URL of the web page which was boiled occasionally, was carried out and URL acquired in many cases, this distinguishes this, and when it is relative description, it changes it into absolute URL. It is as follows if an example is given.

Page URL:. http://www.URL -- data URL:http: [ after .jp/hierarchy1 / abc-01.html data URL:images/momiji.gif conversion ]//www.URL.jp/hierarchy1 / images/momiji.gif[0038]The hour entry acquired in S0308 S0301, URL, the copyright information over a page, From the multimedia data acquired by S0304, and the display name acquired by S0305, SQL for a database register is generated by the SQL (Structured Query Language) generation part 210, and it registers with a database in the database register part 207.

[0039]When confirming whether the investigation of the HTML tag went to the last of a HTML source code and not having gone by S0309 to the last, S0303-S0308 are repeated. For example, it is supposed that it was classified into a category called the result "nature" of having carried out document automatic classification of the HTML file in the HTML file with two or more multimedia data S0303, S0304-S0308 are performed only the number of two or more multimedia data, As a result, when a display name obtains "the little stream of a river", "the mountain in which God lives", "momiji", and two or more attached multimedia data by S0304, the result registered into the multimedia database becomes like drawing 4. The multimedia data said by this example is data of a still picture and an animation, a sound, a document, etc.

[0040]It is possible to build a multimedia database easily from the web page on the Internet by this example by the above, Since correspondence price ****** of the copyright information is carried out with multimedia data, it is possible to perform immediately situation confirmation in the case of utilizing data (the necessity for a right holder's consent, a contact check, etc.).

[0041][Example 2] The difference from Example 1 is having formed a means a user having determined a display name. It is described in detail below. Drawing 5 is a figure explaining the flow of processing of this example. Since S0504 is the same even as S0304 of drawing 3 in drawing 5, explanation here is omitted.

[0042]By S0505, the candidate of a display name for a user to determine a display name is acquired first. It is the following six cases to become a candidate and it acquires them from a HTML source code.

[0043]1. Multimedia data. ALT attribute 2. of the tag for displaying multimedia data. the text

data surrounded by the title tags (<H1 </H1>> etc.) in front of the tag for displaying the text data 4. multimedia data just behind the tag for displaying the text data 3. multimedia data in front of the tag for displaying -- however, When there is two or more multimedia data which makes the same title tag a display name, the last of a display name is numbered.
5. When there are text data surrounded by the title tag (<TITLE> </TITLE>) in the header of the web page, however two or more multimedia data which makes the same title tag a display name, number the last of a display name.
6. What removed extension from multimedia-data-files name by URL of multimedia data [0044] For example, in the case of the HTML source code shown below, the candidate of a display name is as follows.
[0045]
The example of a HTML source code: /nature <TITLE> </HEAD> <BODY> <H1> wild cherry tree of a <HTML> <HEAD> <TITLE> Shiga plateau </H1> ..... Omission ..... The following photograph is taken in the spring of this year. In the spring of <BR><BR><IMGSRC="http://www.URL.jp/hierarchy1/hierarchy2/haru-no-yamazakura.jpg" ALT = " spring of wild cherry tree "> <BR> A wild cherry tree <BR> in full bloom. ..... Omission ..... </BODY> </HTML>[0046]Candidate 1. of a display name In the ALT attribute above-mentioned example of the tag for displaying multimedia data, "the spring wild cherry tree" of the ALT attribute in an IMG tag serves as a candidate.
2. In the text data above-mentioned example in front of the tag for displaying multimedia data, the "spring of this year" in front of an IMG tag becomes a candidate.
3. In the text data above-mentioned example just behind the tag for displaying multimedia data, "the wild cherry tree of spring full bloom" just behind an IMG tag serves as a candidate.
4. In the text data above-mentioned example surrounded by the title tag (<H1> </H1> etc.) in front of the tag for displaying multimedia data, the "wild cherry tree" surrounded by <H1> before an IMG tag </h1> serves as a candidate.
5. In the text data above-mentioned example surrounded by the title tag (<TITLE> </TITLE>) in the header of the web page, "nature of the Shiga plateau" serves as a candidate.
6. With file name which removed protocol, its form (http://), server name, and directory display by URL of multimedia data In the thing above-mentioned example which removed the extension, "haru-no-yamazakura" serves as a candidate.
[0047]Next, the candidate of the display name acquired by S0505 is displayed on the monitor 110 by the message shown in drawing 6, and a display name is determined S0506 because a user chooses a display name.
[0048]Since it is the same as S0306 or subsequent ones of drawing 3 after S0507, explanation is omitted.
[0049]A user is able to choose a display name easily by this example, by the above, when

building a multimedia database from the web page on the Internet.

[0050][Other embodiments] Even if it applies this invention to the system which comprises two or more apparatus (for example, a host computer, an interface device, a reader, a printer, etc.), it may be applied to the device which consists of one apparatus (for example, a copying machine, a facsimile machine).

[0051]So that various kinds of devices may be operated in order to realize the function of an embodiment mentioned above, As opposed to the computer in the device or system connected with these various devices, The program code of the software for realizing the function of the above-mentioned embodiment is supplied, and what was carried out by operating the various above-mentioned devices according to the program stored in the computer (CPU or MPU) of the system or a device is contained under the category of this invention.

[0052]The function of an embodiment which the program code of the above-mentioned software itself mentioned above in this case will be realized, and that program code itself constitutes this invention. the computer network (LAN.) for making program information spread as a subcarrier and supplying it as a transmission medium of the program code The communication media (wire circuits, wireless circuits, etc., such as an optical fiber) in systems, such as WAN, such as the Internet, and a wireless communication network, can be used.

[0053]The recording medium which stored the means for supplying the above-mentioned program code to a computer, for example, this program code, constitutes this invention. As a recording medium which memorizes this program code, a flexible disk, a hard disk, an optical disc, a magneto-optical disc, CD-ROM, magnetic tape, a nonvolatile memory card, ROM, etc. can be used, for example.

[0054]By executing the program code with which the computer was supplied, The function of an above-mentioned embodiment is not only realized, but, Also when the function of an above-mentioned embodiment is realized in collaboration with OS (operating system) or other application software etc. with which the program code is working in a computer, it cannot be overemphasized that this program code is contained in an embodiment of the invention.

[0055]After the supplied program code was stored in the memory with which the function expansion unit connected to the expansion board of a computer or the computer is equipped, Also when the function of an embodiment which CPU etc. with which the expansion board and function expansion unit are equipped based on directions of the program code performed a part or all of actual processing, and mentioned above by the processing is realized, it cannot be overemphasized that it is contained in this invention. Naturally it may be a common computer as shown in drawing 7 as said computer.

[0056]Drawing 7 is a figure showing the internal configuration of common personal installed user terminals. In drawing 7, 1200 is computer PC. . PC1200 was provided with CPU1201 and ROM1202 or the hard disk (HD) 1211 memorized. Or the device control software supplied from

the flexible disk drive (FD) 1212 is performed, and each device connected to the system bath 1204 is controlled in the gross.

[0057]The function of each means of this embodiment is realized by the program memorized by CPU1201 of above-mentioned PC1200, ROM1202, or the hard disk (HD) 1211.

[0058]1203 is RAM and functions as the main memory of CPU1201, a work area, etc. 1205 is a keyboard controller (KBC) and controls the indicating input from the keyboard (KB) 1209, an unillustrated device, etc.

[0059]1206 is a CRT controller (CRTC) and controls the display of CRT display (CRT) 1210. 1207 is a disk controller (DKC) and A boot program (boot program: program which starts the hardware of a personal computer, and soft execution (operation)), plurality -- application -- a compilation file -- a user file -- and -- a network management program -- etc. -- memorizing -- a hard disk -- (-- HD --) -- 1211 -- and -- a flexible disk -- (-- FD --) -- 1212 -- access -- controlling .

[0060]1208 is a Network Interface Card (NIC) and exchanges a network printer, other network equipment, or other PCs and bidirectional data via LAN1220.

[0061]The shape and structure of each part which were shown in the above-mentioned embodiment are only what showed a mere example of the embodiment which hits that each carries out this invention, and the technical scope of this invention must not be restrictively interpreted by these. That is, this invention can be carried out in various forms, without deviating from the pneuma or its main feature.

[0062]

[Effect of the Invention]According to the multimedia database system of this invention, the web page on the Internet, From the multimedia database for using individually to a large-scale multimedia database which is used at a search site, for example, It is possible to register multimedia data into a multimedia database easily, and since copyright information was matched with multimedia data and managed, it is possible to perform immediately situation confirmation in the case of utilizing data (the necessity for a right holder's consent, a contact check, etc.).

---

[Translation done.]